LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JC-153252

Verification and Validation: Goals, Methods, Levels and Metrics

R. W. Logan, C. K. Nitta

April 29, 2003

The Society for Modeling and Simulation International, 2003 Summer Computer Simulation Conference,
Montreal, Quebec, Canada,
July 20-24, 2003

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Verification & Validation: Goals, Methods, Levels, and Metrics

Roger W. Logan
Cynthia K. Nitta
University of California
Lawrence Livermore National Laboratory
PO Box 808 Livermore, CA 94551
rwlogan@llnl.gov
nitta1@llnl.gov

Keywords: Verification, Validation, Uncertainty, Reliability, Confidence

Abstract
This work briefly summarizes the current status of the V&V Program at LLNL regarding goals, methods, timelines, and issues for Verification and Validation (V&V) with Uncertainty Quantification (UQ). Our goals are to evaluate various V&V methods, to apply them to computational simulation analyses, and integrate them into methods for Quantitative Certification techniques for the nuclear stockpile. Methods include qualitative and quantitative V&V processes with numerical values for both (qualitative) V&V Level, and (quantitative) validation statements with confidence-bounded uncertainty bands. We describe the critical nature of high quality analyses with quantified V&V, and the essential role of V&V and UQ at specified Confidence levels in evaluating system certification status. Only with *quantitative validation statements* can rational tradeoffs of various scenarios be made.

INTRODUCTION

It has been said that V&V must address tradeoffs for a "balance of sufficiency and efficiency" (Pilch, 2002), and that V&V must acknowledge (and if we dare, quantify) the point when "better has become the enemy of good enough" as discussed in (Logan and Nitta, 2002). These tradeoffs involve timing and funding for many issues including compute platforms, code development, analyses, and certification issues to be addressed. Part of the planning process assessment of the V&V levels for various certification capabilities. Examples of sensitivity studies, which are part of the prioritization process, are provided. There is a circular dilemma here because we wish to use sensitivity studies to prioritize our V&V efforts, and yet the sensitivity values are only as credible as the V&V we have already done. V&V must therefore be viewed as an evolutionary process in planning, quantification level, and results.

Once we have a working balance of code development, Software Quality, and V&V for specific applications, models with Validation Statements (quantified confidence bounds) of performance and safety margins for various scenarios and issues are applied in assessments of Quantified Reliability at Confidence (QRC). We summarize with a brief description of how these V&V generated QRC quantities fold into a Value-Engineering methodology for evaluating investment strategies. V&V contributes directly to the decision process for investment, through quantification of uncertainties at confidence for margin and reliability assessments.

V&V PROCESS: QUALITATIVE & QUANTITATIVE

To work through the process that leads through V&V and eventually to our investment strategy, it will be useful to track our progress through the methods by referring to the flow process in Figure 1. We break this complex diagram into portions, and discuss each portion in turn.

From Requirements to V&V:

Before we proceed with the V&V process, we have to know the requirements our product or system will have to meet, and which of these our model is to address. After that, depending on both the fiscal and scientific ability to perform a certain level of V&V assessment, we proceed with various degrees of qualitative and (ideally) quantitative validation.

Qualitative Validation:

Although we emphasize our preference for quantitative V&V with numerical confidence bounds, we also recognize the need for *qualitative* recognition of the V&V level for particular simulation capabilities, for example in a "0-10 Meter" numbering system, based on a subset of 84 key criteria identified for a V&V analysis process (Logan and Nitta, 2002).

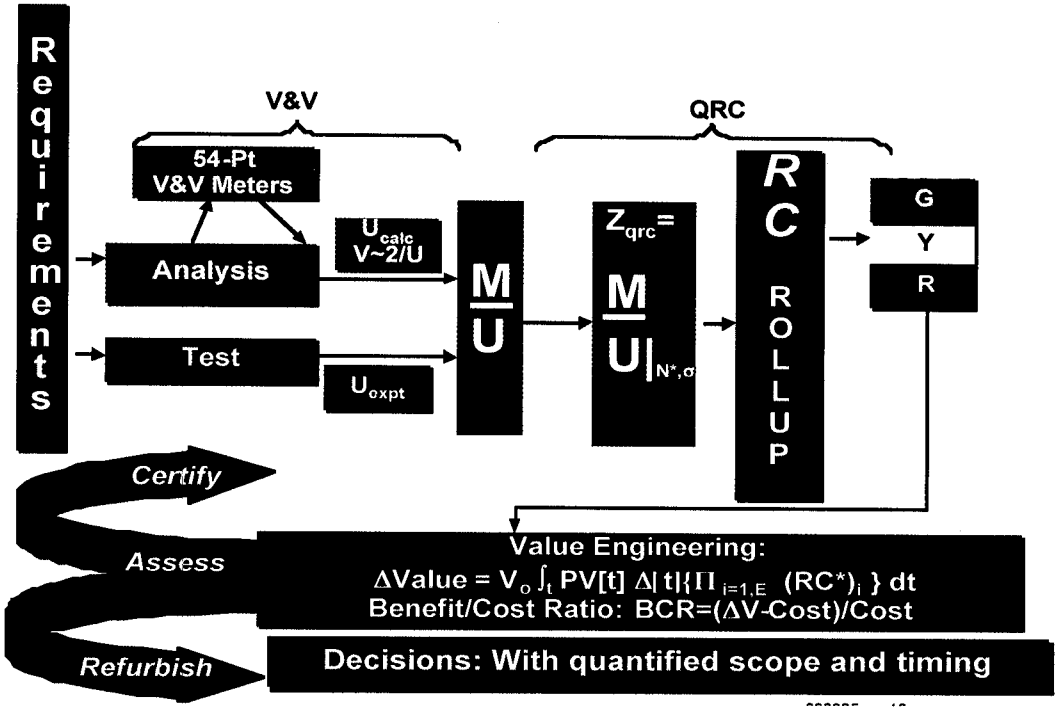

Figure 1. Flow diagram from system Requirements through V&V, through uncertainty quantification and margins; onward through QRC, then QSV (Quantitative System Value) for Value Engineering. The concepts (and acronyms) are easier to grasp as addressed in the full text of (Logan and Nitta, 2002).

For Qualitative Validation, we suggest the use of such a 0-10 rating scale; or for example the Validation Adequacy 0-5 scale from (Trucano et al, 2002). A slightly modified and annotated version of the latter is presented here:

- 0/5="0.0": (Inadequate): No significant comparisons with experimental data – and therefore no measure of correspondence with any such data. These are sometimes very preliminary "what if" analyses. They can be useful as a guide for the next set of analyses, but it is exceedingly dangerous to base any design decisions on them.
- 1/5="0.2": (Inadequate): Ad hoc comparison of experiment "pictures" with prediction "pictures"
- 2/5="0.4": (Incomplete): Ad hoc (nonstatistical) comparisons of experimental data (that may or may not be statistically significant) or data traces
- 3/5="0.6": (Incomplete): Statistical comparison of data and calculations that does not quantify predictive capability of the model or correlation over the parameter space of the database. The degree of extrapolation (if any) may not be quantified. The database may not be statistically significant or fully relevant to the application. *For example, in Figure 2 below, the "database" is reflected in the round data points on the graph. Most of the experimental data falls between [Exhaust] Flow Restriction of 3 and 4. There are certainly enough data points [about 29] to be statistically significant, but we note that all this data was not measured at "standard conditions", e.g. the compression, temperature, etc under which we plan to use the engine. If we take exhaust restriction outside the range of 3 to 4, we are "outside" the parameter space of the database, and we are contending that our validation has gone beyond "Level 3/5=0.6".*
- 4/5="0.8": (Adequate): Predictive capability of the model or correlation is quantified over the parameter space of the database. The degree of *extrapolation* is quantified. *For example, in Figure 2, the solid lines of model prediction clearly go outside the experimental database. We can obtain the degree of extrapolation of exhaust flow restriction directly from Figure 2, but the extrapolation of other engine design quantities from the data is not obvious in Figure 2.* There is a statistically significant database that is fully relevant to the application.
- 5/5="1.0": (Adequate): Predictive capability of the model or correlation is quantified over the full parameter space of the application. *As implied in Figure 2, the application (Flow Restriction range from 0 to 5) may extend well beyond the parameter space of the database (Flow Restriction about 3 to 4).* There is a statistically significant database that is fully relevant to the application. *We can contend that the analysis output shown below in Figure 2 meets this criteria; but that is not obvious from the*

limited information presented in that plot. Even so, quantitative validation of the model is not a statement that the model is adequate; it is only an attempt to supply the information needed for an adequacy assessment.

We suggest that as a first step in the introduction of formal V&V methods to the analysis documentation process, such a rating system, e.g. "0-5", or "0-10", or even "2/5=0.4" be used. An estimate of the V&V level of the work should be stated in the text by the authors of the analysis report. One might first think that authors would tend to be overly optimistic – or pessimistic – in rating their own work. In implementing this qualitative "V&V Level" process, we concede that this does happen, but usually not by much more than 1 level in 10. That is, if the author[s] rate their analysis as "0.45", internal reviews would not tend to rate the same work lower than "0.30" or higher than "0.60". We feel that some such overall rating is a good first step to quantitative V&V for two major reasons:

[a] Some type of overall rating is better than having an upper-level management audience wonder if the work was at "Level 0.10" or "Level 0.95".

[b] The diligence of most analysts, having rated their own work at say "Level 0.4", will make them ask themselves, "how could I make this work reach "Level 0.6, and would that be worth the effort?" This will motivate analysis and documentation to move from *Qualitative* Validation into *Quantitative* Validation.

***Quantitative* Validation:**

For Validation to achieve a *qualitative* rating goal greater than 4 of 10 or greater than 2 of 5 (see examples just above), the work should include the simulations and analysis necessary to generate a Quantitative Validation Statement supporting the *(annotated referring to Figure 2 for this example)* definition of Validation given by (Cafeo and Roache, 2002):

'Validated Model: A model that has confidence bounds on the output *(the middle solid line in the plot)*. A validated model output has the following characteristics:

1. The quantity of interest *(Power Output)*
2. An estimate of the bias *(i.e. confidence bounds are not centered around the model output)*
3. A set of confidence bounds *(the outer solid lines in the plot, drawn at an assessed confidence level)*

A validated model is one where we can support with evidence a formal statement such as:

"I am 90% confident that if I build and measure the quantity of interest, that it will fall within the confidence bands (of uncertainty) shown around the model output."'

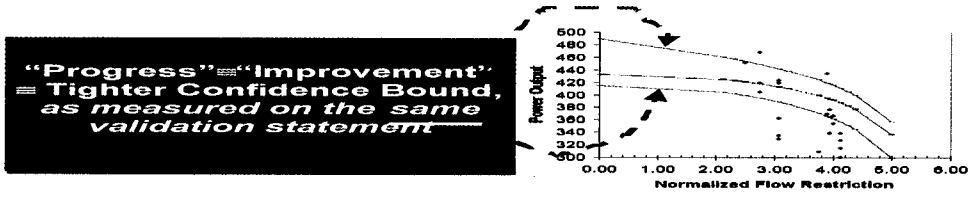

Figure 2. An approach to measure and quantify progress in V&V (and to relate this to dollar value as described below): Quantified measure of progress due to V&V – made possible via use of Quantitative Validation Statements -- can be converted to assessed Risk Reduction and Return On Investment. We stress that improvement lies not in what we can *sketch*, but in what we can *assess* quantitatively.

Figure 2 shows a validation study example from a design application of an internal combustion engine. The plot shows typical output from a Quantitative Validation process. The plot shows numerous small dots of experimental data for this engine design, obtained with various combinations of parameters such as air intake flow, compression ratio for combustion, exhaust flow restriction (the horizontal axis), and other parameters, and the resulting power output of the various tested combinations (vertical axis). If our intended application of the engine requires noise suppression (exhaust flow restriction), we wish to know how much restriction we can use and incur an acceptable reduction in our decision quantity (power output). We develop a model, providing the middle solid line, as a model based answer to answer this question. Quantitative Validation requires a quantitative assessment of the confidence bounds on this model.

This is determined by comparison to a known number of data points (N) as shown in the plot. The model error between the measured output and model output is generally too small to be shown in the graph above, but is used, along with experimental error, variability, and assumed probability distribution functions (PDFs) to construct a confidence bound (the outer solid lines in Figure 2) on our analysis. Any adjustable (calibration) parameters used in the model must be counted as model degrees of freedom (K), so our effective number of data points becomes (N-K). Fewer data points (N) or more model adjustables (K) will result in wider confidence bound lines in the plot. Model adjustables are a fact of life; there is neither the time nor funds to avoid them all. It is simply important to quantitatively account for them in the validation assessment. This simple discussion outlines one such method. A validated model with a Quantitative Validation Statement for performance analyses has the following features, as depicted graphically in Figure 2 and discussed in (Logan and Nitta, 2003):

- The Validation has a statistical nature, as in "coin flipping" analyses, but in our case we must account for the inevitable fact that each "coin" [i.e. test] may be slightly different – and there are often far fewer "coins" than we wish.
- The Validation must provide uncertainty at a stated [and quantified] statistical confidence
- The Validation must show the origins of its data for comparison (the "coins"), and the [perhaps expert judgment] weightings used (are the coins the "same")
- The Validation must allow us to assess a Reliability measure from Margin and assumed or known Probability Distribution Function, normal distribution or other
- The Validation must *provide the information* to address adequacy, before stating whether a given model is "validated for its application" or not
- The Validation must implicitly address the balance of Sufficiency and Efficiency (Pilch, 2002)

V&V and Quantified Reliability at Confidence, QRC

Given a model expressed with this Quantitative Validation Statement and sufficiently quantified information about the system requirements in its environments, we can then assess measures of Quantified Reliability at Confidence (QRC). An overly simplified way to show this using the example in Figure 2 is as follows. If the required power level for our application is "300", and we deploy our design at Flow Restriction of "3", our model assessment is that Power will be about "420". Our Margin "M" of power can be expressed as

M=420-300=120 [1]

If we assume a Gaussian nature for the confidence bounded uncertainties as shown, we can assess our uncertainty in Power as the distance from the model assessed power output (420) and the lower uncertainty bound, the lower solid line, with a Power value of about "390". (The many terms in the uncertainty assessment is beyond the scope of this short discourse). We then have Uncertainty "U":

U=420-390=30 [2]

Our quantitative validation method requires that any such "U" be evaluated at an assessed level of confidence. If we meet this requirement and for the simple example here assume a normal Gaussian distribution for the plot of Figure 2, we can then use the statistical quantity "Z" as

$Z_{qrc}=(M/U)|_{Xs}$ =120/30=4 [3]

Where Xs is the number "X" of Gaussian standard deviations, leading to percent Confidence "C".

The use of Z_{qrc}, taken directly from the statistical "Z", has several advantages. Most important is that it leads to Quantified Reliability "R" at Confidence "C" (QRC). The value of "R" in this simple example is simply the 1-tailed area under the statistical "Z" curve up to the computed value of Z_{qrc} from Equation [3]. The value of "C" is simply the normalized fractional area under our assumed normal distribution at the chosen confidence level Xs, for example "68% Confidence using 1s". *It is vital that we remember that QRC is not the "system reliability". If our model were perfect and the data used in the validation were plentiful with complete relevance, such would be the case. Rather, QRC is a measure of the reliability as our model is credible to assess the quantity. An improvement (or decrement) in QRC may represent a change in the physical system, or simply a change in our model's assessment credibility in that assessment.* When we assess QRC at the system level, we can combine this with measures of Quantitative System Value (QSV) and place these values on a Risk Diagram for investment and decision inputs. The whole process therefore lets us:

- Relate Margin "M" and Uncertainty "U" to "R".
- Demand that we associate "R" with a "C" - and quantifies that "C" based on the specific number of model and test data quantities used for the validation
- Show how better assessments of M and U - and increasing the "effective number of coin flips " N* - quantitatively tightens U, allowing higher quantified C.
- Measure quantitatively our progress in charts like Figure 2.
- Use the quantities in Figure 2 (QRC) as decision inputs via the Risk Diagram in Figure 3.

QRC and Benefit/Cost Ratio, BCR

After the Quantified Reliability at Confidence portion of the analysis, we use these quantities in a model for quantifying business decisions based on inputs from V&V and QRC, with consideration of priority, timing, deployment, and investment strategy. A Quantified Systems Value (QSV) is defined as a function of Reliability and Confidence in terms of Benefit (improvement in Value) and Benefit/Cost Ratios (BCR). For a simple example, consider a product or event with an assessed dollar value QSV0 (see the horizontal axis on the Risk Matrix in Figure 3). If the assessed value of the product is proportional to *its lower-bound model assessed reliability* (expressed as QRC), then the value assessment of the product from our validated model is simply

$$QSV=QSV0*QRC \quad [4]$$

This simple example contains a number of assumptions, for example that there is no undue penalty for the instances (QRC<1) where the model assessment indicates the product does not perform as required. The Benefit "B" of having the product is then this QSV value (our product is assessed to work). We can improve benefit "B" in at least 2 ways; either by improving the physical product (and hence the next assessed QRC), or by improving the model; refining and lowering uncertainty and hence also raising the *assessed* QRC. The latter is important because it allows us to attach a direct dollar benefit DB to the V&V process! Of course either improving the physical product (tighter manufacturing tolerances, etc.) or improving the model (V&V, model or code improvements, etc.) will cost a dollar amount "DC". We can use a Benefit/Cost Ratio:

$$BCR = (DB-DC)/DC \quad [5]$$

The BCR, computed here for product (or model) improvement, gives us a quantity to help answer the question, "was our product or model improvement process worth the cost?" We link the V&V level for particular simulation capabilities (including validation experiments) to the value of products and product decisions made under budget and schedule constraints. A concept of closure is introduced in the form of a simple equation (shown in Figure 1) that integrates UQ, QRC, and QSV quantities with the economic function of Present Value Factor (PV_F) in the time domain. This equation enables quantification of Benefit/Cost tradeoffs and timing decisions. Although there is not a unique BCR, we should explore the bounds of its values for any given decision and we show its relationship to quantified V&V. These concepts are evaluated for particular system requirements, which are in the general case those environments that determine product performance.

Key to the investment strategy process, and its linkage back to V&V, is the Benefit/Cost Ratio (BCR). Quantified V&V shows us that there is not a unique BCR – we must explore its bounds for any given decision. Due to the non-uniqueness of any given BCR, it will become apparent that our decisions fall into 3 basic bins:

1. High BCR within our V&V bounds: Positive decision indicator [i.e. "do it"]
2. Low BCR within our V&V bounds: Negative decision indicator [i.e. "don't do it"]
3. BCR varies high to low depending on V&V bounds: more quantification is needed

CONCLUSIONS: QUANTIFIED VALUE OF V&V

The end product methodology and dollar benefit can be explained using a Risk=Likelihood*Consequence Matrix. "Risk" can also be quantified and viewed, as we will illustrate, as the "Risk" due to inadequate or mistimed V&V. The use of the BCR enables us to balance the benefits of qualitative and quantitative V&V and timing in a demonstrable way. It is obvious that too little V&V is insufficient, while too much V&V is inefficient. The use of a quantified Risk Matrix and the BCR method lets us show how we can determine the level of V&V we feel is appropriate. The evolution from V&V to Reliability at Confidence to Risk is suggested as a tangible way to justify the benefits of investment in V&V.

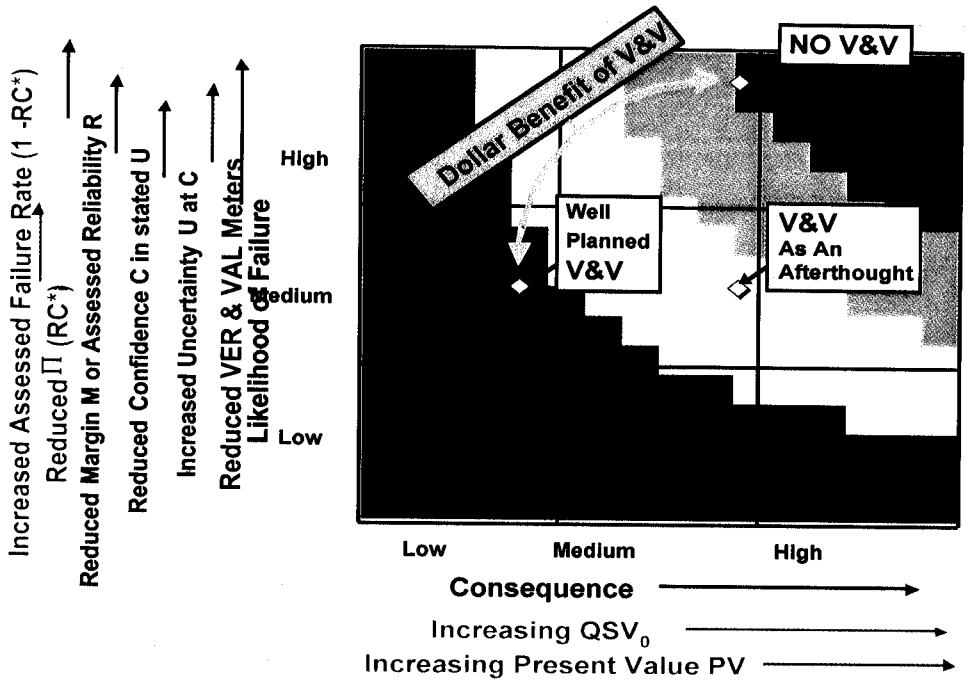

Figure 3. Dollar Benefit of V&V and Quantitative Certification, expressed as a standard Risk=Likelihood*Consequence Matrix. The analogies are built step by step in the full text and in (Logan and Nitta, 2002). Likelihood becomes analogous to assessed (1-QRC); Consequence is expressed in Value Engineering [ie dollars] terms.

REFERENCES:

(Cafeo and Roache, 2002): J.A. Cafeo and P.J. Roache, private communication of draft V&V definitions, April, 2002.

(Logan and Nitta, 2002): R.W. Logan and C.K. Nitta, "Verification & Validation (V&V) Methodology and Quantitative Reliability at Confidence (QRC): Basis for an Investment Strategy", LLNL UCRL-ID-150874, 8 Nov 2002.

(Logan and Nitta, 2003): R.W. Logan and C.K. Nitta, "Validation, Uncertainty, and Quantitative Reliability at Confidence (QRC)", AIAA-2003-1337, Jan 2003.

(Pilch, 2002): M. Pilch, presentation at ASCI-NN Review, 28 October 2002.

(Trucano et al, 2002): T.G. Trucano, M. Pilch, and W.L. Oberkampf, "General Concepts for Experimental Validation of ASCI Code Applications", SAND-2002-0341, March 2002.

Biographies

Dr. Roger W. Logan is Chief Scientist for Weaponization at LLNL. He has led program elements and technical efforts in weapon system assessment, simulation, and experimentation. He has over 20 years experience in code development, simulation, and more recently, in formal mathematical methods for V&V from a user standpoint and from a benefit/cost investment perspective. He has written extensively on the use of simulation for weapon system certification during the current era of extended periods of time in the absence of nuclear testing. He holds B.S., M.S., and Ph.D. degrees from University of Michigan, University of California, and University of Michigan respectively.

Dr. Cynthia K. Nitta is currently LLNL's Verification & Validation (V&V) Program Leader and has held numerous positions in the technical and leadership aspects of nuclear design and certification. She has extensive experience ranging from code development to simulation to integration of V&V into the weapon assessment and certification process. She is a 2003 LLNL Edward Teller Fellow and is a design physics group leader in the Defense and Nuclear Technologies Directorate. She holds a B.S. from Princeton University with M.S. and Sc.D. degrees from the Massachusetts Institute of Technology.

University of California
Lawrence Livermore National Laboratory
Technical Information Department
Livermore, CA 94551